УДК 004.62

А. К. Бызова, С. Л. Гольдштейн

ОЦЕНКА ТОЧНОСТИ КЛАССИФИКАЦИИ ТЕКСТОВ В ЗАВИСИМОСТИ ОТ ИХ ЧИСЛА СРЕДСТВАМИ DATA MINING

Рассматривается проблема достижения оптимальной точности классификации вербальных текстов средствами Data Mining. Эмпирически оценена точность классификации в зависимости от числа обучающих текстов и количества классов. Также автором рассматривается зависимость точности классификации от представления исходных данных для обучения классификатора: в виде научных статей и в виде словарей терминов. В работе приведены результаты обучения и тестирования классификатора при различных факторах, описанных выше. В качестве средства Data Mining выбрана программа Weka.

Ключевые слова: *дерево принятий решений, интеллектуальный анализ данных, классификация текстов, компьютерная лингвистика, обучение классификатора, Weka.*

The problem of optimal classification accuracy of verbal texts by means of Data Mining. Empirically evaluated the classification accuracy depending on the number of training texts and the number of classes. The author examines the dependence of the classification accuracy of the representation of the original data for training the classifier: in the form of scientific articles and a glossary of terms. The results of training and testing of the classifier for various factors described above. As a means of Data Mining program is selected Weka.

Keywords: *decision tree, data mining, text classification, computational linguistics, supervised learning, Weka.*

В виду непрерывного роста объема информации возникает проблема ее структуризации в целях быстрого нахождения и извлечения из этого множества необходимых сведений. Использование инструментов Data Mining позволяет находить неочевидные закономерности — знания. Исходя из совокупности наиболее часто появляющихся слов в тексте можно предсказать, к какой предметной области относится данный текст. При этом полнота и точность — характеристики, по которым можно определить качество такой классификации. Известно, что полнота и точность классификации текстов — меры, противоречащие друг другу, т. е. 100 %-ную полноту легко достичь, поместив все тексты в *i*-й класс, при этом точность будет мала. И, наоборот, 100 %-ную точность легко достичь, строго отбрасывая тексты, помещая в *i*-й класс малое число текстов — полнота будет мала [1]. Так, например, достаточными для классификации по индивидуальному стилю будут тексты объемом в 800 предложений или 9 000 слов (минимум 6 000) [2]. Также ранее нами было установлено, что с ростом количества классов возникает более высокая степень неопределенности, а значит, для достижения точности классификации технических текстов более 50 % требуется большее число текстов для построения модели [3].

Универсальной зависимости точности классификации текстов от числа текстов и количества классов нет. Поэтому в данной статье поставлена и решена задача эмпирическим путем установить зависимость точности классификации текстов от их количества и числа классов.

Обоснование выбора программного пакета и текстов

В качестве Data Mining-средства выбран программный пакет Weka [4]. Для проведения эксперимента тексты для обучения были взяты из базы данных научных статей [5], они представляют собой англоязычные тексты объемом в одну страницу. Также нами заранее был подготовлен словарь англоязычных существительных по 1 000 терминов для каждого класса.

Под мерой точности будем понимать отношение количества правильно классифицированных текстов к числу неправильно классифицированных.

Обучение

Процесс интеллектуального анализа текстов реализован нами поэтапно:

1. Предварительная обработка: отобранные тексты импортируем в формат *.arff (Attribute-Relation File Format).
2. Обучение, начиная с исходных данных, но с одним текстом из технического класса, затем с двумя и т. д. до десяти (в качестве метода выбран алгоритм дерева решений).
3. Проверка: испытание модели на тестовой выборке.

На первом этапе произвели ввод текстов для обучения по классам. Количество классов заранее известно. Далее происходит работа со словами. Пусть $C = \{c_i\}$ — множество классов, $T = \{t_j\}$ — обучающее множество, которое содержит термины, каждый из которых характеризуется атрибутами, при этом один из них указывает на принадлежность к классу. Логику построения дерева можно свести к следующим продукционным правилам:

if T содержит хотя бы один термин, относящихся к классу C_i
  then следующий узел для T — это класс C_i;
if T не содержит ни одного из терминов
  then следующий узел для T — это класс, ассоциированный с родителем;
if T содержит термины, относящиеся к разным классам
  then T разбиваем на подмножества.

Для разбиения выбирается один из признаков, имеющий два и более отличных друг от друга значений $O_1, O_2, \ldots O_n$. Требуется: распознать принадлежность каждого из 10 (15) текстов к классу технических; провести серию из 4-х экспериментов в вариациях для 2-х, 4-х, 8-ми и 16-ти классов. Множество T разбивается на подмножества $T_1, T_2, \ldots T_n$, где каждое подмножество T_i содержит все примеры, имеющие значение O_i для выбранного признака. Это процедура рекурсивно продолжается до тех пор, пока конечное множество не будет состоять из примеров, относящихся к одному и тому же классу. В итоге получаем дерево.

Распознавание

Теперь остается проверить адекватность дерева решений на тестовых текстах.

Эксперимент 1

Имеется два класса текстов (химические и технические). Исходные данные для построения модели отобраны в виде 20 (затем 30) научных англоязычных текстов объемом около 3 000 слов (*.txt), 10 (15) для технических текстов и 10 (15) для химических текстов. В качестве тестовых данных — 10 (15) технических текстов такого же объема на том же языке (*.txt). Требуется: распознать принадлежность каждого из 10 (15) текстов к классу технических наук; провести серию из 4-х экспериментов в вариациях для 2-х, 4-х, 8-ми и 16-ти классов.

Для 1-го варианта построили 10 моделей. На каждой испытали тестовую выборку. Результаты тестирования представлены на рис. 1а.

Для 2-го варианта увеличили количество классов вдвое, добавив два класса: экономические, медицинские. Исходные данные при построении модели были отобраны в виде 40 научных текстов для каждого из классов. Результаты приведены на рис. 1б.

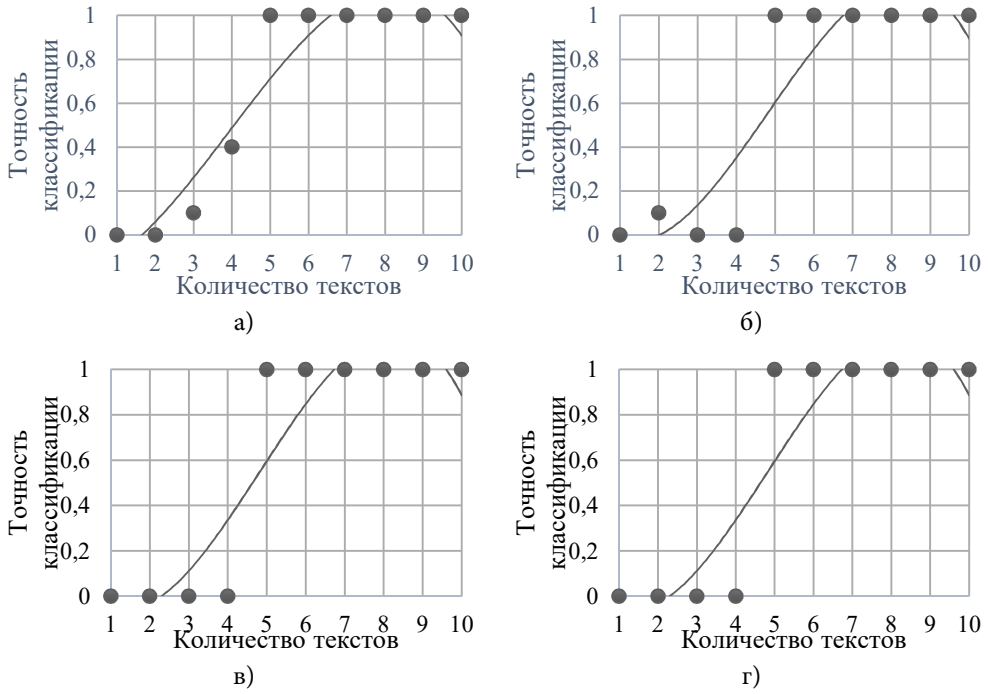

Рис. 1. Зависимость точности классификации от числа текстов для 2-х (а), 4-х (б), 8-ми (в) и 16-ти (г) классов при 10-ти текстах в каждом из классов

Для 3-го варианта увеличили количество классов вдвое, добавив 4 класса: юридические, физико-математические, искусствоведческие, психологические. Исходные данные были отобраны в виде 80 научных текстов для каждого из классов.

Результаты приведены на рис. 1в. Для 4-го варианта увеличили количество классов еще раз вдвое, добавив 8 классов: биологические, исторические, науки о Земле, педагогические, политологические, сельскохозяйственные, социологические, филологические. Исходные данные были отобраны в виде 160 научных текстов для каждого из классов. Результаты приведены на рис. 1г.

Аналогично, для 15 текстов для 1-го варианта получили 15 моделей. На каждой испытали тестовую выборку. Результаты тестирования представлены на рис. 2а.

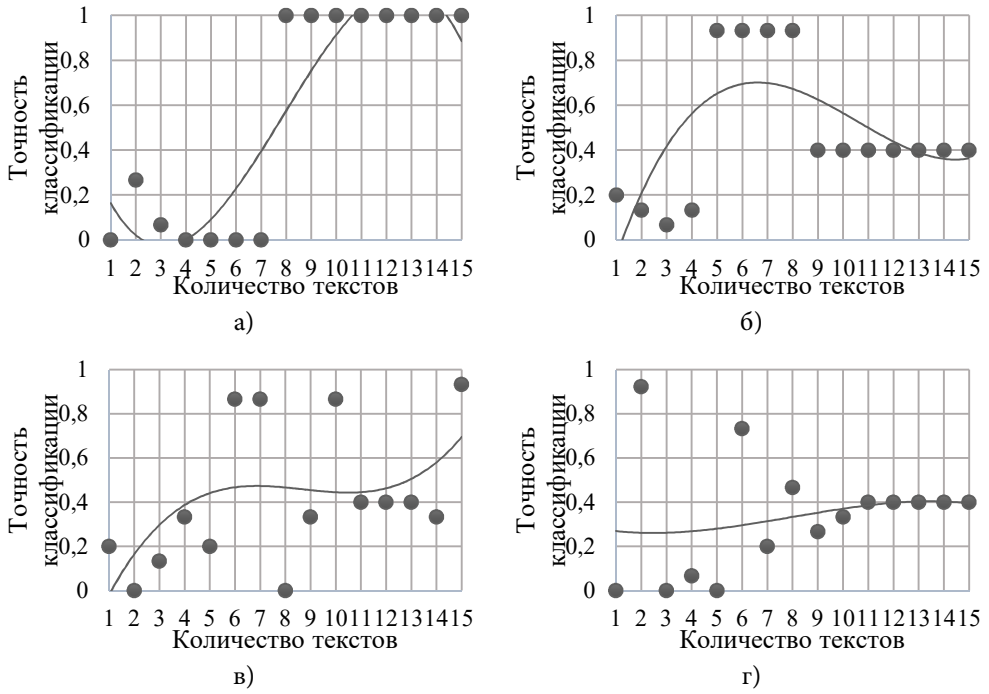

Рис. 2. Зависимость точности классификации от числа текстов для 2-х (а), 4-х (б), 8-ми (в) и 16-ти (г) классов при 15-ти текстах в каждом из классов

Для 2-го варианта исходные данные при построении модели отобраны в виде 60 научных текстов для каждого из классов (рис. 2б). Для 3-го варианта — в виде 120 научных текстов для каждого из классов (рис. 2в). Для 4-го варианта — в виде 240 научных текстов для каждого из классов (рис. 2г).

Поскольку классов стало больше, возникает более высокая степень неопределенности. При этом существуют тексты, содержание которых может комбинировать в себе несколько научных направлений. Это также приводит к неопределенности.

Сильный разброс точек можно объяснить тем, что при построении моделей мы отбрасывали те тексты, которые содержали в себе большую часть ключевых терминов. Таким образом, для достижения оптимальной точности классификации следует обращать внимание на выборку исходных данных. Возникает потребность в создании универсальной онтологии научных терминов.

Эксперимент 2

Имеется два класса текстов (химические и технические). Исходные данные для построения модели отобраны в виде двух словарей для каждого класса. Требуется: распознать принадлежность научной статьи к классу технических наук; провести серию из 3-х экспериментов в вариациях для 2-х, 4-х и 6-ти классов. Результаты приведены в табл. 1–4. В каждом словаре — 1 000 терминов, дробление словаря происходит на равные части, «+» — текст распознался, «–» — текст не распознался.

Таблица 1

Результаты распознавания для 2 классов

Кол-во текстов для обучения	Результаты распознавания по классам*	
	c (химия)	e (экономика)
e1 c1	–	+
e2 c2	–	+
e5 c5	–	+
e10 c10	+	+

Таблица 2

Результаты распознавания для 3 классов

Кол-во текстов для обучения	Результаты распознавания по классам*		
	c	e	l
e1 c1 l1	+	–(c)	–(c)
e2 c2 l2	–(l)	–(l)	+
e5 c5 l5	–(l)	–(l)	–(e)
e10 c10 l10	–(l)	+	+

Таблица 3

Результаты распознавания для 4 классов

Кол-во текстов для обучения	Результаты распознавания по классам*			
	c	e	l	w
e1 c1 l1 w1	–(e)	+	–(e)	–(e)
e2 c2 l2 w2	–(l)	–(l)	+	–(l)
e5 c5 l5 w5	–(l)	–(l)	–(e)	–(l)
e10 c10 l10 w10	–(l)	+	+	+

Таблица 4

Результаты распознавания для 5 классов

Кол-во текстов для обучения	Результаты распознавания по классам*				
	c	e	l	w	m
e1 c1 l1 w1 m1	–(w)	–(w)	–(w)	+	–(w)
e2 c2 l2 w2 m2	–(l)	–(l)	+	–(l)	–(l)
e5 c5 l5 w5 m5	–(w)	+	+	+	–(w)
e10 c10 l10 w10 m10	–(m)	+	+	+	–(l)

* c — химия, e — экономика, l — лингвистика, w — юриспруденция, m — машиностроение

Видно, что при увеличении количества классов необходимо большее число текстов. За счет тщательной подготовки материала можно снизить количество текстов для обучения.

Результаты и выводы:

– поставлена задача эмпирическим путем установить зависимость точности классификации текстов от их количества и числа классов,

– в качестве Data Mining средства выбрана программа Weka,

– проведены два эксперимента,

– получена зависимость точности классификации текстов от их числа и от количества классов,

– рассмотрены два варианта выбора исходных данных: в виде научных статей и в виде словарей терминов.

Программный пакет Weka пригоден для классификации англоязычных текстов. Однако следует уделять большее внимание на отбору исходного материала для обучения. Если в качестве исходных данных использовать словарь терминов, то сам процесс классификации упрощается.

Литература

1. *Большакова Е. И., Клышинский Э. С., Ландэ Д. В. и др.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 272 с.
2. *Шевелев О. Г., Петраков А. В.* Классификация текстов с помощью деревьев решений и сетей прямого распространения // Вестник Том. гос. ун-та. 2006. № 290. С. 300–307.
3. *Бызова А. К. Гольдштейн С. Л.* Оценка точности классификации текстов в зависимости от их числа средствами Data Mining // Тезисы докладов II Междунар. молодежн. научн. конф.: Физика. Технологии. Инновации ФТИ-2015 (20–24 апреля 2015 г.). Екатеринбург: УрФУ, 2015. С. 117–119.
4. Computer Science Department, University of Waikato. URL: http://www.cs.waikato.ac.nz
5. Архив научных статей [Электронный ресурс]. URL: http://www.gramota.net/materials.html (дата обращения: 24.03.2015).